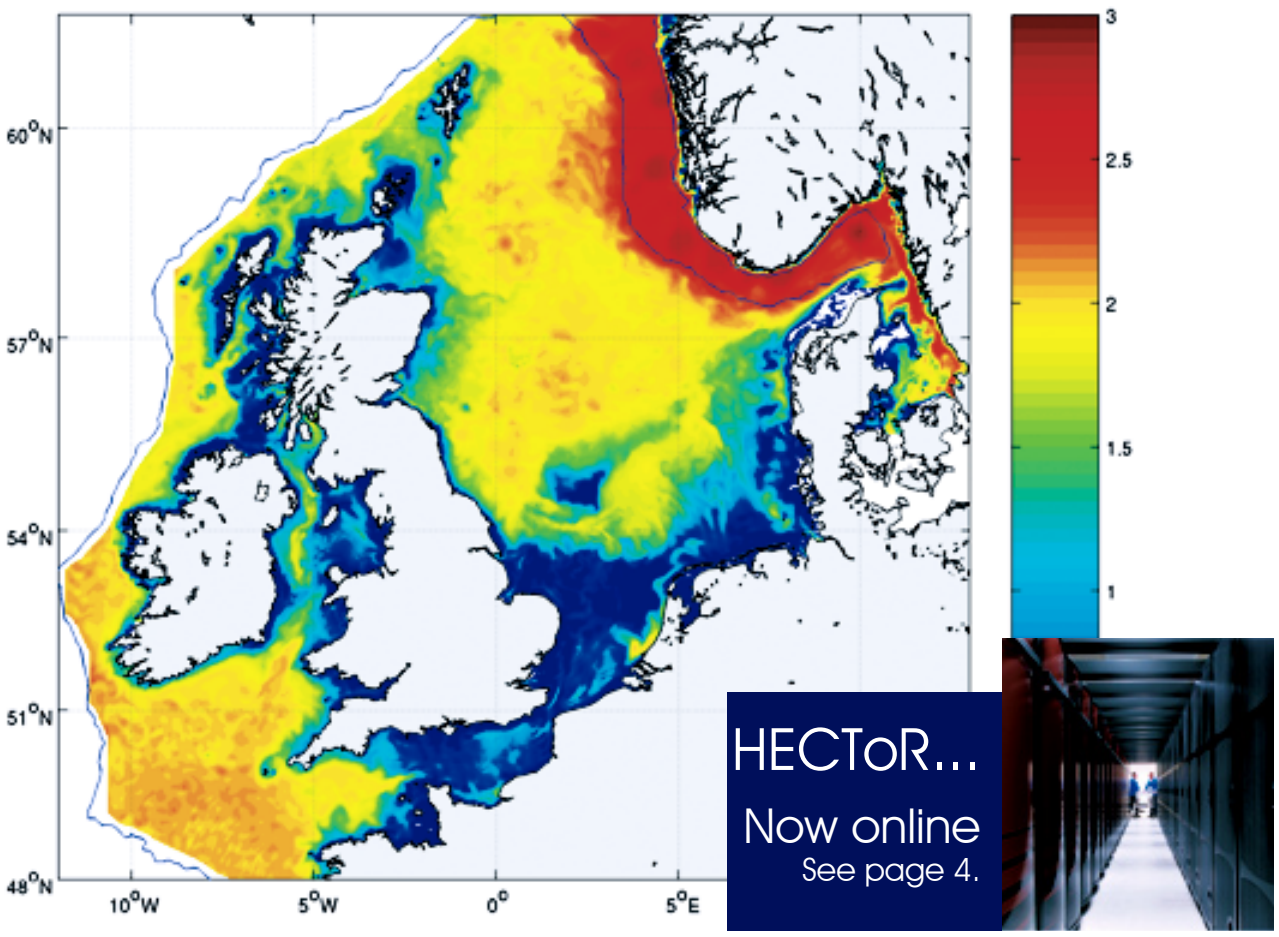
Capability Computing

The newsletter of the HPCx community

[ISSUE 10, AUTUMN 2007]



HECToR...
Now online
See page 4.

Modelling the ocean with HPCx

CONTENTS

Editorial

Kenton D'Mellow, EPCC

Welcome to the latest edition of *Capability Computing*, the HPCx community newsletter. This landmark tenth edition marks five successful years of service, over which the machine itself has taken many forms. We are now well into Phase 3, and have recently opened a new large jobs queue of 1536 processors. We hope to interest several consortia with this prospect.

These are exciting times for the HPC community. In this issue we announce the arrival of HECToR, the next generation of UK national academic supercomputing service. The HECToR machine is located at the University of Edinburgh, and the service is brought to you by the same organisation that runs HPCx (UoE HPCX Ltd), in partnership with Cray Supercomputers Inc. and NAG (Numerical Algorithms Group). We also aim to open a dialogue on the future of HPCx in partnership with HECToR – we believe this to be a unique opportunity to provide complementary HPC services to the UK academic community.

The main focus is, as always, on enabling science. This edition includes feature articles on state-of-the-art oceanographic simulations, scientific calculations enabled by novel computational methods that specifically exploit parallelism, the modelling of radiation damage in metals, and charge and energy transfer processes in nanoscale systems.

We are also pleased to present technical updates and primers from our own Applications Support and Terascaling teams: these include single-node performance of the Phase 3 system, a guide to profiling tools and techniques, and the brand new service administration software (SAFE).

Finally, it is with great sadness that we report the death of John Fisher, our Head of User Support. John made a great contribution to the success of HPCx and will be sadly missed. An appreciation can be found on page 5.

I hope you enjoy this edition, and hope you can attend the HPCx Annual Seminar in November. See the back page for further details.

Sixth DEISA training event

Gavin J. Pringle, EPCC

Image courtesy of HLRS.

As regular readers of Capability Computing will know, HPCx is a member of the Distributed European Infrastructure for Supercomputing Applications, or DEISA.

DEISA seeks to deploy and operate a persistent, production quality, distributed supercomputing environment with continental scope, and to enable scientific discovery across a broad spectrum of science and technology. Scientific impact (enabling new science) is the only criterion for success.

The sixth DEISA Training Session will be held in Stuttgart, Germany, in February 2008. Registration for this event will open early in the New Year. Scientists from all European countries and members of industrial organisations involved in high performance computing are invited to attend.

The purpose of the training is to enable fast development of user skills and the knowledge needed for the efficient utilisation of the DEISA infrastructure.

The first part of the training event will give a global description and introduction to the DEISA infrastructure and will describe the general middleware services, the use of the DEISA Common Production Environment and the detailed utilisation of UNICORE and the DESHL (a command line interface and API to UNICORE). The second part of the training will focus on a particular aspect of HPC programming of general interest to DEISA users. As in the past, this session will also include talks from scientists who have successfully exploited DEISA for their research purposes.

Most academic attendees currently based outwith Germany will have their travel expenses reimbursed.

For more information, please visit http://www.deisa.eu/training/

Complementary capability computing and the future of HPCx

Alan D Simpson
EPCC Technical Director
Director of UoE HPCX Ltd

Although EPSRC has had a long-standing policy of having two overlapping national HPC services, different funding models and management structures have made collaboration between these services challenging. However, UoE HPCX Ltd, a wholly-owned subsidiary of the University of Edinburgh, now has a leading role in both HPCx and HECToR. This is a unique opportunity for these services to work together to maximise the benefits for UK research.

Both services are very much focused on the delivery of Capability Computing, and although HECToR is more powerful, HPCx still represents a significant fraction of the UK's total HPC capability. There are clear benefits to the user community in providing different hardware systems with different strengths and weaknesses. For example, while HECToR allows scaling to a larger number of processors, HPCx's shared memory supports large memory jobs.

However, the benefits can be increased by the way the services are configured. When HECToR becomes the premiere national service by the end of 2007, there will be more flexibility in the set up of HPCx, particularly allowing a wider variety of job classes. Certain job classes, such as very long jobs or pre-booked slots, are likely to have an adverse impact on overall utilisation on the prime national service but could be accommodated on a complementary service.

Over the last few months, we have been discussing with EPSRC complementary capability roles for HPCx to maximise cthe ombined research output. To ensure that user requirements are central, EPSRC organised a workshop with a selection of users from a broad range of application areas. The workshop demonstrated strong support for the concept and we are planning a trial on HPCx towards the end of this year that will include:

- Simple mechanisms for transferring resources between HPCx and HECToR
- Advance booking on HPCx
- 48-hour jobs on HPCx
- 15-minute debugging jobs on HPCx.

There was also strong support for:

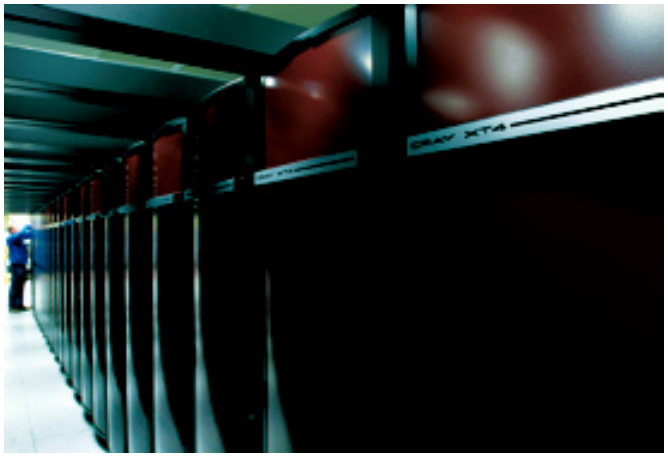
- Ensemble computing
- Large memory jobs (utilising SMP nodes)
- Data-intensive jobs – especially for NERC users.

We are convinced that our central role in both the HPCx and HECToR services is an ideal opportunity to demonstrate the advantages of overlapping and complementary HPC services. While we have already identified a number of complementary capability roles for HPCx, we anticipate that more will emerge as users seek to maximise the benefits from the two services.

We invite users to give us their views on the complementary role of HPCx by completing the survey available from the SAFE user login main page (https://www.hpcx.ac.uk/) or by emailing any specific requests to: helpdesk@hpcx.ac.uk

HECToR service is ready ahead of schedule!

John Fisher, EPCC User Support
Alan D Simpson, EPCC Technical Director



'I am delighted to say that we rose to the challenge of installing the HECToR service in record time and, in parallel with the user support acceptance testing, passed all of the service provision acceptance tests with flying colours. The cost of failure for us and the UK computational science community would have been very high and our success is due in no small part to the efforts of EPCC staff and our colleagues at Daresbury, Cray and NAG.' Prof. Arthur Trew, HECToR Service Director

The HECToR service has officially passed all its acceptance tests and formally starts in October 2007. After the long procurement process for this service, the final few months have gone extremely well and HECToR has entered service ahead of schedule. This is due to the efforts of a great many people at Cray and at EPCC, including Mike Brown and his team, Stephen Booth and John Fisher.

HECToR (High-End Computing Terascale Resource) is the UK's new high-end computing resource, funded by the Research Councils and available primarily for academics at UK universities as well as researchers throughout Europe. The first phase of HECToR is a Cray XT4-based system with a peak performance of around 60 Tflops, and is located at the University of Edinburgh's Advanced Computing Facility (ACF). The contracts were signed in February and it is remarkable that the significant upgrades required to the ACF have been completed successfully to what was a very aggressive timetable. Indeed, early users were on the system in the middle of September, less than 7 months after contract signature!

The University of Edinburgh, through its wholly-owned subsidiary, UoE HPCX Ltd, holds the prime contract for service provision, including technology, accommodation and management, and the helpdesk. This work will be subcontracted to Cray, STFC's Daresbury Laboratory and EPCC. EPCC and Daresbury Laboratory also provide HPCx, the UK's existing high-performance computing service. NAG Ltd will provide the computational science and engineering support for HECToR.

EPCC's role

EPCC will host all of the HECToR hardware at the University of Edinburgh's recently upgraded ACF building. Through Mike Brown, EPCC leads the joint EPCC-Daresbury Operations and Systems Group which is responsible for running the HECToR systems and associated infrastructure. EPCC also provides the User Support and Liaison team, which is responsible for the helpdesk, the website and third party application codes. Much of the administration of the service will be conducted on-line through the SAFE (Service Administration from EPCC) which was developed by Stephen Booth.

ACF facts and figures

The main HECToR systems will be accommodated in a newly-renovated computer room of around 300m^2. However, the new plant room, which provides power and cooling, is even bigger at around 450m^2.

HECToR Phase 1

The main part of the initial HECToR Phase 1 configuration is a scalar Cray XT4 system, plus associated storage systems.

The Cray XT4 is contained in 60 cabinets and comprises 1416 compute blades, each of which has 4 dual-core processor sockets. This amounts to a total of 11,328 cores, each of which acts as a single CPU. The processor is an AMD 2.8 GHz Opteron. Each dual-core socket shares 6 GB of memory, giving a total of 33.2 TB. The theoretical peak performance of the system is over 60 Tflops, with a LINPACK performance in excess of 52 Tflops.

There are 24 service blades, each with 2 dual-core processor sockets. They act as login nodes, as controllers for I/O and for the network. Each dual-core socket controls a Cray SeaStar2™ routing and communications chip. This has 6 links which can implement a 3D-torus of processors. The point-to-point bandwidth is more than 2 GB/s, and the minimum bi-section bandwidth is over 4 TB/s. The latency between two nodes is around 6 µs.

The storage solution consists of some 40 TB of home space, which will be regularly backed up, plus over 500 TB of high-performance work space which uses the Lustre distributed parallel file system.

Future upgrades

The scalar Phase 1 system will be supplemented during 2008 by a Cray vector system known as 'BlackWidow'. This will include 28 vector compute nodes; each node has 4 Cray vector processors,

Continued opposite.

John Fisher

1945–2007

I am very sad to announce that John Fisher who was Head of EPCC User Support, died suddenly on Thursday 27th September 2007. His death will sadden all those who knew him throughout the UK computational science community but most especially those of us who had the privilege to work closely with him.

EPCC was a significant part of John's working life and John was a significant part of EPCC. John joined EPCC as Head of User Support in 1994 just shortly after I had started. We have worked closely together throughout this time and have worked in adjacent offices for well over a decade. From the beginning, it was clear that he was an interesting individual who would bring his own character to the job.

When he joined EPCC, John already had a successful career in computing spanning almost a quarter of a century, working at such companies as 3L and Lattice Logic. Nevertheless, taking on the User Support role at a supercomputing centre was a major challenge. I'm sure it is no surprise to those who knew him to hear that he fully rose to this new opportunity. John's time at EPCC was a success both for him as an individual and for the centre. His first role was supporting the UK's first parallel supercomputing service on a Cray T3D, and he subsequently took on a similar role in the HPCx service.

John had been planning to retire in a year or two at the end of the HPCx service, but had been a major contributor to our successful bid for HECToR, which is based on a next-generation Cray system. John had put in sterling efforts over the last few months to ensure that HECToR was ready for service and he was happy and relieved when the service passed its acceptance tests less than a week before he died. I know he was justifiably proud of his contributions to this.

In all of these supercomputing services, John's role was communicating with the hundreds of users from throughout the UK. This was a role that he was ideally suited for. His comforting presence and sympathetic ear made him the friendly face of supercomputing support in the UK for more than a dozen years. He moulded this role and made it his own. He will be impossible to replace.

Many people have commented on his humour and wit; his kindness and friendliness; his erudition; his calm and professional approach. John was a great communicator and could always find the right way to smooth over a problem or an unhappy user. Many people at EPCC will fondly remember the beautiful way he expressed his wit in his humorous emails on the occasion of his 60th birthday or when looking for volunteers to cover queries over Christmas. He brought light to our working lives.

John was a well-liked and highly respected colleague who made enormous contributions to the success of EPCC. But he was more than that. John was our friend; John was my friend. It is the little things that I will miss: being able to pop into his office at the end of a difficult day for a quick chat and a few kindly words; being able to rely on his comforting presence; his passionate and cheery nature.

I had the great privilege and pleasure of working with John for this first part of my own working life. He has been a great support and influence on me and on the whole of EPCC. EPCC may never be the same place again but I know that I have benefited enormously from knowing him, and that his presence will continue to influence and guide us.

I still find it hard to believe that someone who was so full of life is no longer with us and I keep expecting him to pop in with some humorous story to tell. The world is a sadder, greyer place without John in it but it was a privilege to know him and I am proud to call him my friend.

Alan Simpson, Technical Director, EPCC.

If you wish to leave your own condolence message, we have set up a webpage at: http://www.epcc.ed.ac.uk/jf/

---

making 112 processors in all. Each processor is capable of 20 Gflops, giving a peak performance of more than 2 Tflops. Each 4-processor node shares 32 GB of memory. The storage capacity will also be almost doubled during Phase 1.

The next stage of the project, planned for October 2009, will take the peak performance up to 250 Tflops. This may be a mixture of next-generation scalar systems, based on quad-core Opterons, and BlackWidow vector systems; the balance will be determined following a review of users' requirements. A third phase is planned for 2011.

Outlook

The last few months have been an incredibly busy and challenging period. We are delighted to be ready to offer a user service ahead of schedule and are very much looking forward to the next 6 years. HECToR will be an excellent addition to the UK's computational resources and should ensure that UK computational research remains world-class.

To find out more about the HECToR service, including how to apply for time, see: http://www.hector.ac.uk/

Single node performance on HPCx Phase 3

J. Mark Bull, EPCC

In order to understand the single-node performance of applications on Phase 3 of the HPCx system, we used hardware counters to instrument a set of 19 codes, most of which consume a significant number of cycles on the service. We found that most codes run at between 8% and 20% of the nominal peak floating point performance of the system. A small number of codes, which heavily utilize tuned libraries, run at between 20% and 50% of peak. We also investigated the performance impact of enabling simultaneous multithreading (SMT): each application was also run using double the number of processes, but on the same number of physical processors with SMT enabled. The performance gain varied from a 29% slowdown to a 44% speedup.

In addition to measuring the floating point performance, we also use the hardware counter facilities of the POWER5 processor to record a range of other metrics for each code.

In this study, we used parallel versions of all the codes, and collected data from runs using all 16 processors of a p575 shared memory node. The 19 application codes used were:

Quantum Chemistry: AIMPRO, CASTEP, CRYSTAL, GAMESS-UK, SIESTA, VASP

Molecular Dynamics: DL_POLY, LAMMPS, MDCASK, NAMD, AMBER-PMEMD, AMBER-SANDER

Computational Fluid Dynamics: NEWT, PCHAN, LUDWIG

Atomic Physics: H2MOL, PRMAT

Plasma Physics: CENTORI

Ocean Modeling: POLCOMS

Metric	Min	Mean	Max	Correlation
Flops per FP load/store (CI)	0.72	2.00	4.18	0.88
Floating point load/stores per cycle	0.08	0.30	0.49	0.79
Instructions per nanosecond	0.71	1.74	2.68	0.77
Percentage of instructions which are FP	21.61	54.75	88.42	0.75
Percentage of load/stores which are FP	33.48	63.98	91.60	0.66
Percentage of flops in FMAs	17.47	73.04	99.69	0.66
Level 1 cache hits per nanosecond	0.23	0.61	0.93	0.65
Level 1 cache references per nanosecond	0.28	0.66	0.95	0.61
Level 1 cache hit ratio	81.15	90.99	98.00	0.44
TLB misses per microsecond	0.03	0.27	0.80	0.04
Level 3 cache misses per microsecond	0.01	0.42	3.11	-0.02
Memory accesses per microsecond	0.004	0.39	3.11	-0.03
Level 3 cache hits per microsecond	0.09	1.24	4.44	-0.09
Level 2 cache hit ratio	91.27	96.65	99.19	-0.11
Level 3 cache hit ratio	8.72	78.23	98.36	-0.12
Level 2 cache hits per microsecond	17.81	53.99	131.17	-0.13

Table 1: Minimum, average and maximum values of performance metrics, and correlation coefficients of performance metrics with flop rate.

hpmcount

hpmcount is a utility which gives access to the hardware counters on the IBM POWER5 processor. There are six hardware counter registers on the processor: two of which always count the same events: processor run cycles and completed instructions. The other four registers can count a wide variety of possible events, many of which are not especially useful for performance analysis. In this study we used the AIX version of hpmcount, which only allows access to eight groups of events. The ACTC version, which was not available at the time, but which is now installed on HPCx, gives access to 148 event groups.

Methodology

Each benchmark or application was run on 16 processors (with a few exceptions, see below). For each code, five separate runs were made, instrumented with hpmcount and recording events from counter groups 1, 2, 3, 4, and 8. (Note: for reasons related to symmetry in the problem geometry, H2MOL can only execute on processor numbers of the form n(n+1)/2. Therefore instead of 16 processors, runs were made on 15 processors. In addition, the NEWT and PRMAT runs were made on 32 processors (2 nodes) and the AIMPRO runs on 64 processors (4 nodes), due to restriction on problem size and geometry. In these cases, the data were scaled to single node appropriately.)

For each application (except GAMESS-UK, which failed to execute correctly), we performed a second set of runs where the number of processes (i.e. MPI tasks) was doubled (or increased to 28 in the case of H2MOL) and SMT enabled, with the number of physical processors used remaining the same.

Floating point performance

For each code, the total flop rate for all 16 processors was calculated, and expressed as a percentage of the peak flop rate for the p575, which is 96 Gflop/s. This data is shown in Figure 1.

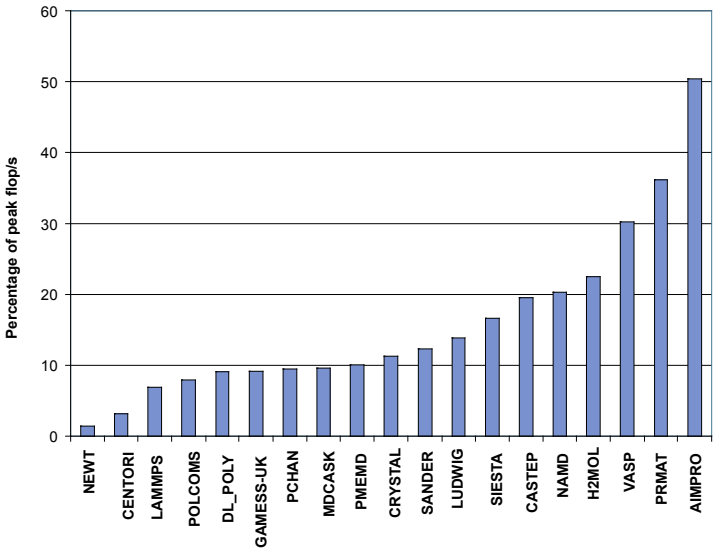


Figure 1: Percentage of peak performance achieved.

The highest performance (just over 50% of peak) is obtained by AIMPRO and the lowest, 1.5% of peak, by NEWT. A small group of codes achieve between 20% and 50% of peak performance. The codes in this group typically make extensive use of tuned libraries (either dense linear algebra or FFTs) for a significant portion of the

computation. Most other codes achieve between 8% and 20% of peak performance.

It must be noted that the percentage of peak performance is not necessarily a good measure of the quality of an application, and should not be interpreted as such. Flop rates are a poor measure the quality of science achieved, and can vary significantly within a single application depending on the input data and chosen parameters. Furthermore, some types of application are intrinsically harder for modern microprocessors to achieve high flop rates on than others.

Other metrics

In addition to the flop rate, a number of other metrics can be derived from the raw hardware event counts. In order to assess the importance of these metrics in determining the flop rate, we computed the correlation coefficient between each metric and the flop rate across the 19 codes. These correlation coefficients are shown in Table 1, together with the minimum, mean and maximum values of the metrics across the 19 codes. For this size of dataset (19 variables, 17 degrees of freedom), a correlation coefficient with absolute value of greater than 0.456 is significant at the 95% level, and a correlation coefficient with absolute value of greater than 0.575 is significant at the 99% level.

The metrics with the highest correlation to flop rate are flops per floating point load/store (called computational intensity in hpmcount output) and floating point load stores per cycle. This is not surprising, since the flop rate can be written as the product of flops per floating point load/store with floating point load/store per cycle. Figure 2 shows how these two metrics combine to give the resulting flop rates.

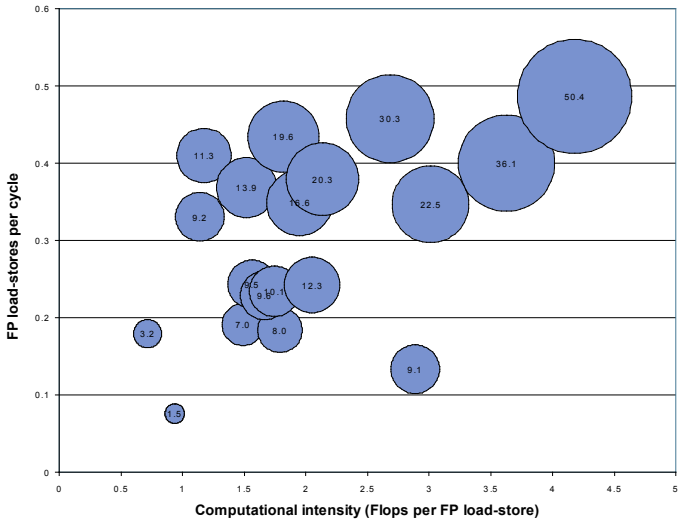


Figure 2: Scatter plot of floating point load/stores per cycle against computational intensity. The size of each circle and the number it contains is the percentage of peak flop rate.

Codes which obtain a high percentage of the peak flop rate have a high computational intensity (between 2.5 and 4) and also perform a high number of floating point load/stores per cycle (between 0.3 and 0.5). Many of the other codes have a computational intensity of between 1 and 2.5, and these appear to fall into two groups, one with a high number of floating point load/stores per cycle (between 0.3 and 0.5) and one with a lower number of floating point load/stores per cycle (between 0.15 and 0.3). Two codes (NEWT and CENTORI) have a computational intensity of less than 1. DL_POLY is alone in having a high computational intensity, but a low floating point load/store rate.

Simultaneous multithreading

For each application, we compared the total execution time for the runs with and without SMT enabled. Figure 3 shows the speedup for each application as a result of enabling SMT: a number less than one indicates that the code ran slower with SMT than without. Since the benchmark cases were chosen such that communication is not a significant overhead, we can be reasonably confident that the increased communication overheads resulting from doubling the number of processes are not important in determining the outcome.

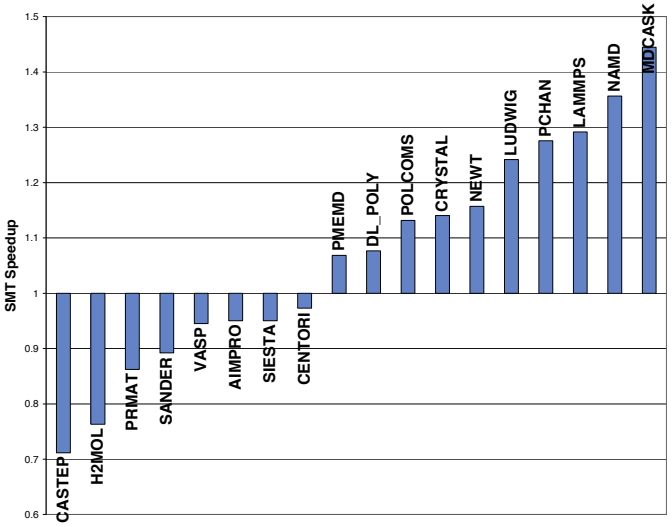


Figure 3: Speedup resulting from enabling SMT. The results show a range of outcomes, from a 29% slowdown for CASTEP to a 44% speedup for MDCASK. The geometric mean speedup is 1.06 and the geometric mean speedup of applications which benefit from SMT is 1.22.

Further Reading

Additional results, including analysis of some of the other metrics can be found in the HPCx technical report 'Single Node Performance Analysis of Applications on HPCx' available at http://www.hpcx.ac.uk/research/hpc/technical_reports/HPCxTR0703.pdf

Acknowledgements

The author gratefully acknowledges the assistance of the following people in collecting data for this study: Martin Plummer, Ian Bush, Andy Sunderland, and Mike Ashworth of STFC Daresbury Laboratory; Alan Gray, Joachim Hein, Jon Hill, Kenton D'Mellow, Fiona Reid, Lorna Smith and Kevin Stratford of EPCC, The University of Edinburgh.

High-resolution modelling of the Northwest European shelf seas using POLCOMS

Jason Holt and Roger Proctor, Proudman Oceanographic Laboratory, Liverpool, UK
Mike Ashworth, STFC Daresbury Laboratory, Warrington, UK

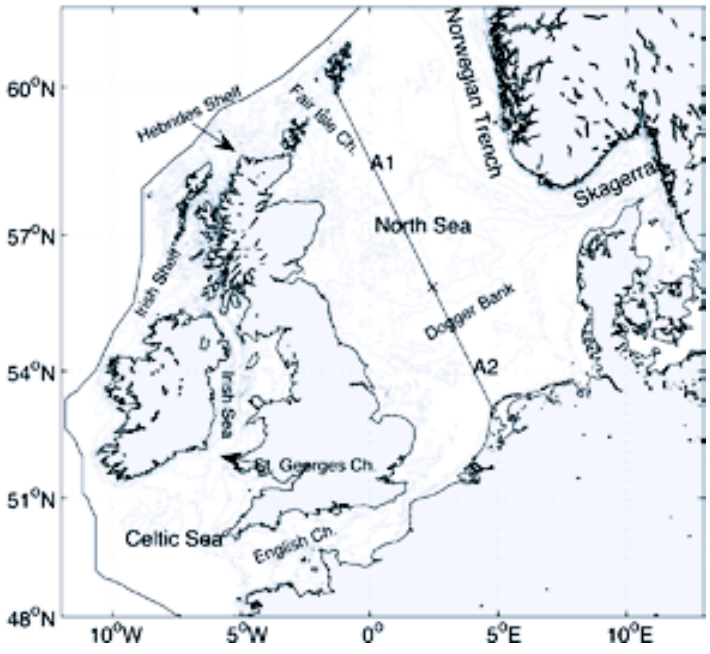


Figure 1: The model domain showing the nesting (left) of this model (HRCS) within the Atlantic Margin Model (AMM). The dashed line shows the 200m isobath. The panel on the right shows the model bathymetry: contour interval is 20m up to 100m and thereafter100m.

The POLCOMS High Resolution Continental Shelf model is the finest resolution model to-date to simulate the circulation, temperature and salinity of the Northwest European continental shelf. It has enough resolution (about 1 nautical mile) to include the important small-scale, density-driven features and enough coverage to include the large-scale circulation across the shelf.

The Proudman Oceanographic Laboratory Coastal Ocean Modelling System (POLCOMS) has been used to simulate the seasonal cycle for 2001, and with a series of model experiments has been able to separate the model currents into components driven by the winds, by changes in density, and those driven by the North Atlantic (including tides). It was previously thought that the winds dominated the long term circulation for most of the year, but these results demonstrate the importance of the density circulation for a much greater portion of the year and over a wider area than previously expected. What is the reason for this? The density currents might be small, but, like the net tidal transport, they always act in the same direction. The winds, on the other hand, might drive stronger currents but are much more variable in direction. The results are important for our understanding of the transport of nutrients, pollutants and dissolved carbon around shelf seas, and have implications for the ecosystem approach to marine management, and for the role of shelf seas in the carbon cycle.

The circulation of the Northwest European continental shelf (Figure 1) has been the subject of intense investigation, because of its importance to the major north European fisheries. The earliest studies focused on the tide and wind driven circulation, but it has long been known that, just as in the deep ocean, horizontal density gradients play a significant role. The sub-tidal circulation on shelf seas is made up of a combination of tidal residuals, wind (and atmospheric pressure) driven currents, and density-driven currents. These forcing mechanisms all have local and non-local manifestations: the imposed stresses and pressure gradients are local effects, while the propagation of coastally-trapped waves is the non-local (free) response to this forcing. In the context of regional scale modelling, the oceanic boundary conditions imposed on a model can form a significant part of the non-local forcing. The aim of this work is to investigate the relative importance of these components for the shelf wide space- and seasonal time-scale transport across the Northwest European shelf.

The characteristic scale of the density structure is the internal Rossby radius, which is typically 2-5 km on the shelf and about 14 km in the Norwegian Trench. Based on these and other important scales, the model resolution for this study is chosen to be ~1.8 km. This represents a compromise between the need to represent small scale features and the need for a practical model size for seasonal integrations. Hence, on the shelf, this model is 'eddy permitting' rather than 'eddy resolving' in the sense that motions at the Rossby radius are included but not well resolved. An eddy resolving model would require a resolution of finer than 200m and would draw in to question the hydrostatic approximation applied in this model. Figure 2 shows the potential energy anomaly (the energy required to fully mix the water column when it is density stratified) during the summer of 2001. It demonstrates the range of small scale features (e.g. eddies and filaments) and sharp fronts reproduced by this model.

Model studies are ideal for investigating the composition of the circulation, since the wind/tide/density contribution cannot generally be established unambiguously from observations.

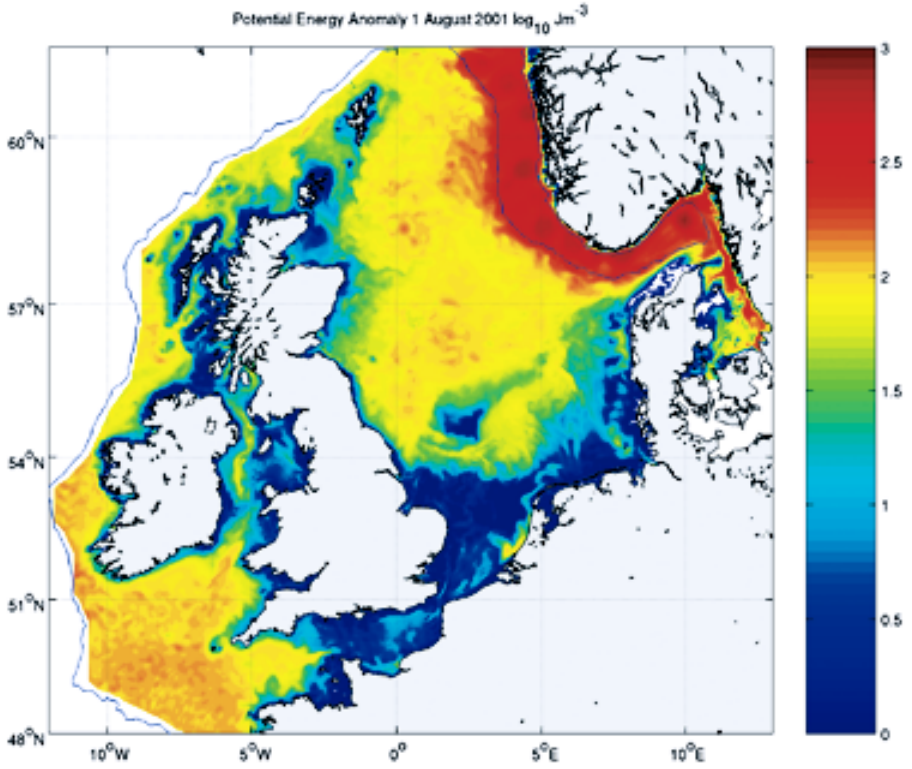


Figure 2. The potential energy anomaly during mid-summer. This is a measure of summertime stratification and shows how much energy would be needed to mix the water column.

A number of baroclinic modelling systems have been used to investigate the circulation of this region. These typically have resolutions of 7-20 km and are often used as the basis for shelf sea operational oceanography. These modelling studies generally agree with the expected circulation pattern on the Northwest European shelf derived from observational studies, but while the general patterns have been reproduced, direct comparisons between model currents and contemporary observations are rare and have not yet been carried out on a shelf-wide scale.

To date there are no published model studies of the density driven circulation across the whole of the Northwest European shelf region that include motions at the scale of the internal Rossby radius, i.e. are eddy permitting. Here we employ a high resolution shelf wide application of POLCOMS in a study of an annual cycle. Unlike in the previous model studies, the resolution considered here is fine enough to resolve the details of the density driven frontal circulation, while the scale is large enough to include all the frontal systems on the shelf and also the large scale circulation.

We use POLCOMS to simulate the region shown in Figure 1 on a 1/60° latitude by 1/40° longitude grid (~1.8 km), with 34 vertical s-coordinate levels. This configuration is known as the High Resolution Continental Shelf (HRCS) model. POLCOMS is a B-grid finite difference model with prognostic temperature and salinity, and a sophisticated advection scheme (the Piecewise Parabolic Method). The simulation period is from 1st January to 31st December 2001 and a multi-year (1998-2001) simulation of the POLCOMS Atlantic Margin Model at ~12 km resolution (AMM in Figure 1) provides initial conditions and contemporary boundary data for HRCS (this in turn is forced by the North Atlantic FOAM Model). Runs were carried out on HPCx using 256 processors. Details of the optimisation of POLCOMS for large-scale parallel platforms are given in [1].

We investigate three one-year long model experiments; 1) full baroclinic, 2) full barotropic and 3) barotropic with only tidal/ oceanic forcing. By sequentially subtracting the results of these experiments we can divide the currents into density, wind and oceanic components. Dividing the currents in this fashion does not separate out the non-linear effects, so is only an approximate indication of the relative importance of the terms and none of these components can be treated as truly independent of the others.

As an example of the model output, the sub-tidal currents (calculated as 25hr means) averaged for July to September and over the depth interval 20m to 40m are shown in Figure 3. Detailed results and a full discussion are presented elsewhere [2]. This high resolution model shows that the circulation across the shelf is made of many small scale structures; the largest currents appear as filament-like jets, with high spatial coherence.

The results show that the HRCS model transports agree well with the available literature estimates of volume flux and, while these estimates are from sparse measurements, from different periods and not exactly co-located with the model sections, they give a semi-quantitative validation of the model results, particularly that it reproduces the relative strengths of currents reasonably well. Previous model results have been compared with extensive observations from the 1988-89 North Sea Project and it is our intention to repeat this comprehensive validation exercise with HRCS in the near future.

These results demonstrate the diversity of structure, and in time and space scales of shelf sea currents, which is only realizable through model simulations that include motions at the scale of the Rossby radius. Wind, density and oceanic forcing are found to play an approximately equal role in the overall circulation pattern in these shelf seas. The importance in the northern North Sea of the density field driving current during the break down of stratification in late summer and autumn is demonstrated.

As the wind increases and convective mixing is introduced during

Continued on page 11.

Profiling parallel codes on HPCx

Andrew Sunderland
STFC Daresbury Laboratory

HPCx users who wish to understand better the parallel performance of their codes should consider using one of the parallel profilers available on HPCx. Both the Vampir[1] and Paraver[2] profiling tools enable detailed graphical representations of parallel application code performance. The performance analysis data can be used to identify such issues as computational and communication bottlenecks, load imbalances and inefficient CPU utilisation. Tracefiles for MPI, OpenMP and mixed-mode MPI/OpenMP can all be generated and analysed using these tools on HPCx. In this article, I focus on the usage and features of Vampir. Recently, new versions of Vampir and VampirTrace[3] have been released which include support for the generic Open Tracefile Format (OTF). This enables users to view VampirTrace output from their personal preference from a selection of compatible tracefile analysers such as Kojak[4] or Tau[5]. Hardware performance counter monitoring using the Performance Application Programming Interface (PAPI) is also now supported in the latest releases of Vampir and VampirTrace.

Both analysis tools involve similar approaches, ie analysis of a specific tracefile created at the application's runtime that contains information on the various calls and events undertaken. Tracefile generation using VampirTrace requires a relinking of the application code to the VT library before running the parallel code in the usual way. By default the whole parallel run is traced, but if desired, tracing can be switched on and off within the source code itself through calls to the VampirTrace API. Meanwhile, environment variable settings can be used to customize the tracing events that are to be recorded. The Vampir analysis tool can then used to convert the trace information into a variety of graphical views, e.g. timeline displays showing state changes and communication, profiling statistics displaying the execution times of routines, communication statistics indicating volumes and transmission rates, and more.

A typical view from Vampir is shown in Figure 1. This view of the Activity Chart trace gives a breakdown of the time taken in MPI communication by routine.

A more detailed output for understanding an application's parallel characteristics is the timeline view, shown in Figure 2. Here time is represented along the horizontal axis with the processes listed vertically. Time spent in computation is represented in green and time spent in communication in red. Message passing between processes is represented by the black (point-to-point) and purple (global communication operations) lines that link the process timelines. This timeline represents a common pattern for many application codes, where a series of time-steps in a calculation involve repeated phases of computation separated by phases of global communication.

Zooming in further, the characteristic communication pattern for a 3 dimensional Fast Fourier Transform (FFT) can be displayed (Figure 3a) i.e. pairwise point-to-point communications in firstly the x, then the y, then the z directions. A new feature in Vampir 5 is that a corresponding counter timeline showing performance metrics can also be displayed. Figure 3b shows how the Flops/s rate reduces to almost zero during communication-dominated periods and serial performance peaks at around 100 Mflops/s during this FFT computation.

Figure 1: Vampir Activity Chart for specific MPI routines.

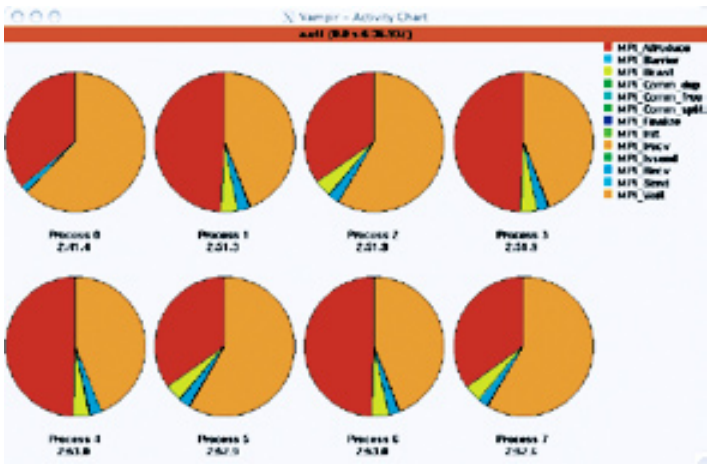


Figure 2: Extract from Global Timeline View showing coarse-scale communication pattern.

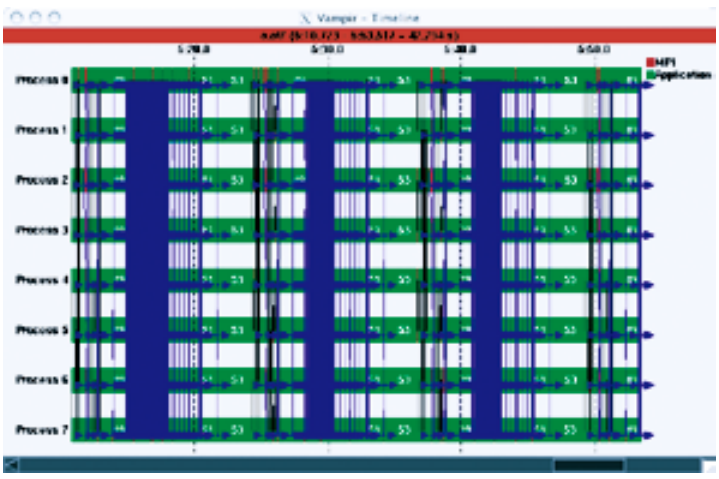
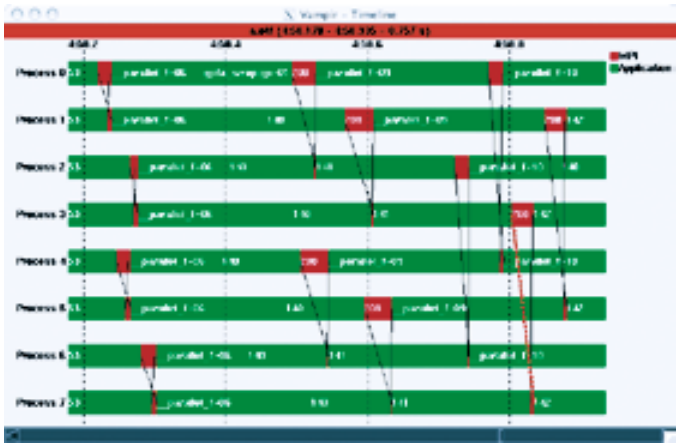
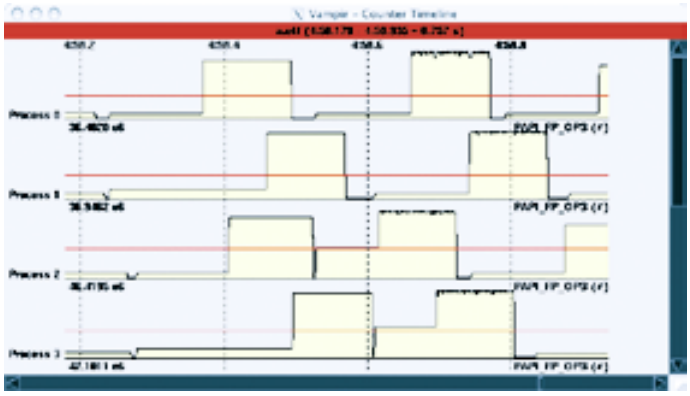


Figure 3a and 3b: Fine-grained communication pattern for 3D FFT with associated performance metric timeline.





Further details about both Vampir and Paraver usage on HPCx can be obtained from the HPCx User Guide[6] and from a recently published HPCx Technical Report[7].

References

[1] Vampir – Performance Optimization http://www.vampir.eu

[2] Paraver, The European Center for Parallelism of Barcelona, http://www.cepba.upc.es/paraver

[3] Vampirtrace, ZIH, Technische Universitat, Dresden, http://tu-dresden.de/die_tu_dresden/zentrale_einrichtungen/zih

[4] KOJAK – Automatic Performance Analysis Toolset, Forschungszentrum Julich, http://www.fz-juelich.de/zam/kojak/

[5] TAU – Tuning and Analysis Utilities, University of Oregon, http://www.cs.uoregon.edu/research/tau/home.php

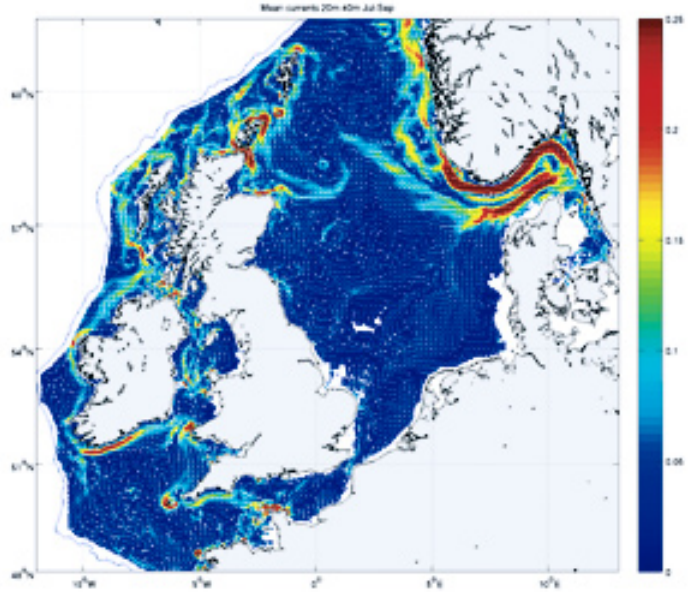
[6] User's Guide to the HPCx Service, http://www.hpcx.ac.uk/support/documentation/UserGuide/HPCxuser/HPCxuser.html

[7] Profiling Parallel Performance using Vampir and Paraver, HPCx Technical Report HPCxTR0704, http://www.hpcx.ac.uk/research/hpc/technical_reports/HPCxTR0704.pdf

High-resolution modelling continued

the autumn, these fronts, and their corresponding currents migrate across the shelf and persist over a number of months: in many regions the stratification is not completely broken down until December. These results suggest that the effects of the density driven circulation are more wide-ranging both in terms of time and space than the fixed location of tidal mixing fronts might suggest. The surface and benthic boundary layers play an important role in determining the density field throughout the year and these results demonstrate that the accurate modelling of vertical mixing is crucially important to the modelling of shelf sea circulation. This work has not addressed the critical issue of lateral mixing between coastal waters and water of Atlantic origin, which determines the overall exchange of material between the land and open-ocean. Such a study requires more than a single annual cycle to be simulated, since the flushing time of the North Sea is a substantial fraction of a year. Such a study is the subject of on going work.

Figure 3. Summer mean horizontal circulation averaged at mid depth (20-40m). For clarity only velocity vectors at every 8th grid point are shown. The colour shading indicates current speeds and is from the full model results.



[1] M Ashworth, JT Holt, R Proctor, "Optimization of the POLCOMS hydrodynamic code for terascale high-performance computers", Proceedings of the 18th International Parallel & Distributed Processing Symposium, 26th-30th April 2004, Santa Fe, New Mexico, 2004.

[2] JT Holt and R Proctor, "The seasonal circulation and volume transport on the Northwest European Shelf: a fine resolution study", Journal of Geophysical Research, 2007, in press.

Modelling non-adiabatic processes in materials with correlated electron-ion dynamics

Daniel Dundas and Eunan J McEniry, School of Mathematics and Physics, Queen's University Belfast
Daniel Mason, Department of Physics, Imperial College London
Lorenzo Stella, Department of Physics and Astronomy, University College London and The London Centre for Nanotechnology

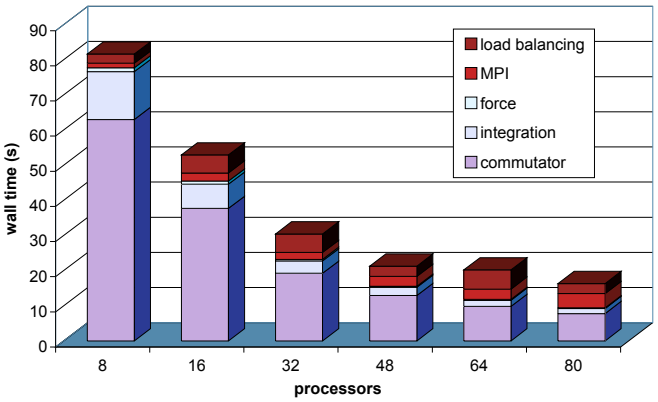


Figure 1 (far left): Performance of splCED on HPCx – wallclock time per update step for a 17,473 atom simulation.

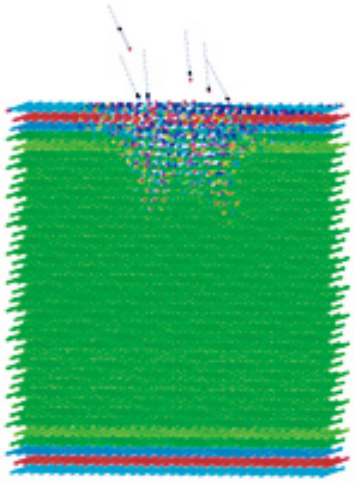


Figure 2 (left): Sputtering simulation using splCED. The system comprises 17473 copper atoms making a film 8nm thick in the y-direction and periodic in the x- and z-directions. A copper ion having a kinetic energy of 500eV is fired at normal incidence into the film. The figure corresponds to a snapshot taken 150fs after the start of the simulation.

An understanding of charge and energy transfer processes is crucial in modelling nanoscale systems in contact with an environment because of their large surface to volume ratio. In modelling these processes, different levels of approximation can be invoked depending upon the problem under investigation. These approximations range from considering the atomic configuration to be essentially fixed in space – the Born-Oppenheimer approximation (BOA) – through mixed quantum-classical descriptions in which a quantum evolution of the electronic subsystem is coupled to a classical description of the ionic subsystem – such as the Ehrenfest approximation – to fully quantum descriptions. The principle drawback of making the BOA is that processes involving the irreversible flow of energy cannot be treated. Examples of such irreversible processes include the heating of metallic atomic wires due to current flow; polaron formation in polymer chains; and friction, plastic deformation and radiation damage in metals. While Ehrenfest dynamics does allow energy transfer, it has recently been shown [1] that although a correct description of the heating of cold electrons by energetic ions is provided, the reverse process – namely the heating of ionic vibrations by hot electrons – is wholly incorrect. The reason for this failure of the Ehrenfest approximation (the treatment of the electrons as a structureless fluid together with a neglect of the correlations between quantum fluctuations of the positions and momenta of the ions with those of the electrons) has recently been addressed in a new method called Correlated Electron-Ion Dynamics (CEID) [2].

As part of EPSRC's Materials Modelling Initiative, a number of groups (based at QUB, UCL and Imperial College) have joined together in a consortium to tackle three exemplar areas of fundamental and strategic significance in materials science where irreversible exchanges of energy between electrons and ions play a central role. They are a) inelastic transport in nanostructures; b) radiation damage in metals; and c) excited states in polymers. This collaboration has resulted in the development of theory, algorithms and associated computer codes that exploit the power of high-performance computer architectures, allowing the treatment of progressively more complex systems at increasing levels of sophistication. Two computational packages have been developed within the consortium to address these scientific problems.

splCED: an Ehrenfest dynamics code for radiation damage

At Imperial College, the focus of research within the CEID consortium has been the investigation of energy transfer from the ionic to the electronic subsystems using dynamical simulations within the Ehrenfest approximation. The aim is to understand the role and characteristics of electronic damping in radiation damage and sputtering. The magnitude of this damping has an important influence on the number of Frenkel defects produced in low energy cascades, and can be directly compared to the frictional damping terms used in classical molecular dynamics simulations [3].

Ehrenfest dynamics simultaneously solves Newton's equations for the motion of classical ions with the quantum Liouville equation for the electrons, with an electronic Hamiltonian that depends parametrically on the ionic positions. As each ion has only n interacting neighbours, the Hamiltonian is sparse. The electronic density matrix describes the electronic state, which in a metal is a dense NxN Hermitian matrix, N being the number of tight-binding (TB) orbitals used. This density matrix is not truncated and its memory cost is thus $N^2/2$ complex numbers, which must be distributed across P processors. The main algorithmic cost is the time evolution of the density matrix, evaluated as the commutator

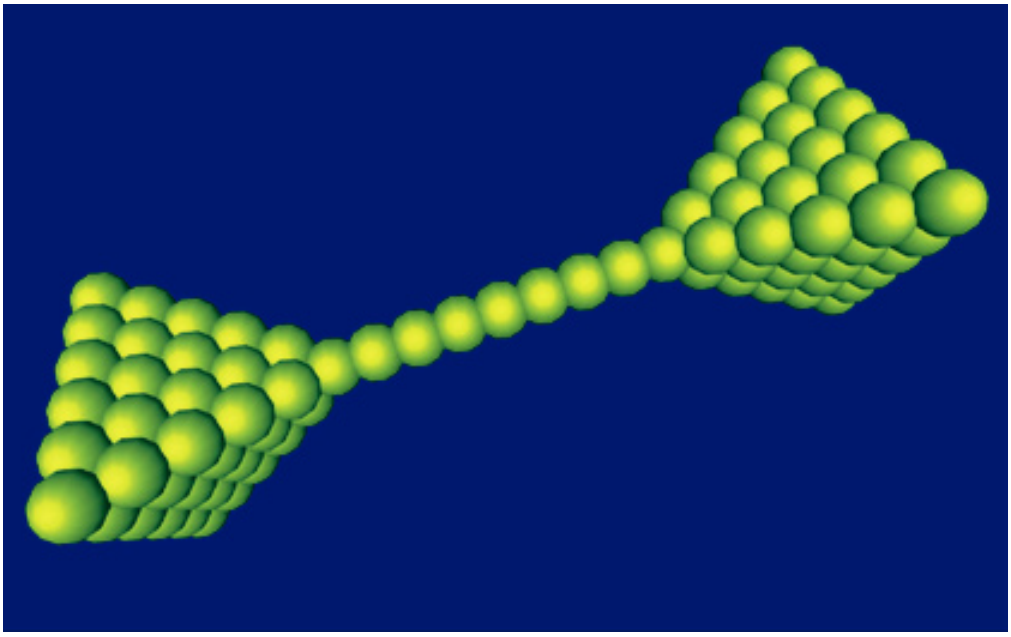


Figure 3: Schematic representation of a nanoscale wire (nine atoms long) connected through two pyramidal contacts to semi-infinite electrodes (the electrodes are not shown). In a typical simulation the electrodes are connected to a charge reservoir (for instance a battery), which results in current flow in the system.

of the Hamiltonian and density matrices, which takes nN^2 floating point operations. A parallel time-dependent TB code (spICED) has been developed to perform Ehrenfest dynamic simulations for systems comprising more than 10,000 atoms and for picosecond timescales. For the first time, we can perform non-adiabatic dynamics in metals at scales hitherto accessible only to classical molecular dynamics. Currently, spICED can handle over 30,000 atoms and scales well up to 128 processors. We still have some load balancing issues, but as the density matrix is updated using only local message passing, we achieve a calculation wall time speed of $3.5nN^2$ nanoseconds per timestep (the timestep is set to 0.05fs simulation time) on 64 processors of HPCx. Figure 1 presents scaling results for a typical simulation for a system comprising 17,473 copper atoms making a film 8nm thick in the y-direction and periodic in the x- and z-directions.

In Figure 2 we present a snapshot taken 150fs after the start of a simulation for the system considered in Figure 1. A copper ion having a kinetic energy of 500eV is fired at normal incidence into the film. In this case we see that six ions are ejected from the surface, with directions indicated in the figure. Atoms are coloured according to their electronic charge, with red denoting 5% electronic charge (ie positive) and blue −5% electronic charge (ie negative). Friedel charge oscillations are seen at the surface, and long focussed collision sequences drive the disturbance deep into the layer.

pDINAMO: an implementation of the CEID equations of motion

The CEID equations extend the Ehrenfest approximation through an expansion of the equations of motion (EOMs) in terms of a hierarchy of electron-ion correlation functions. pDINAMO has evolved from a serial code (DINAMO) originally developed inhouse at QUB to solve the CEID equations. The equations to be solved constitute a tightly coupled set of equations that are currently truncated at second order in electron-ion correlations. This results in a closed set of equations that are propagated in time. As compared to the Ehrenfest approximation, algorithmic and memory costs of the second-order CEID equations scale as $(6M + 1)N^2$ where M is the number of moving CEID ions, and N is the number of basis functions used in our description of the electronic subsystem (currently tight-binding models are used).

One re-engineering phase of pDINAMO has already taken place and a second is commencing. Each of these phases has been initiated by a need to both improve performance and scalability of the code (thus increasing the range of problems that can be tackled) and also in order to implement additions to the continuously developing theory. In order to produce a fully scalable code, two levels of parallelization are required: firstly, the distribution of correlation matrices (the N^2 scaling); secondly, the distribution of correlation matrices corresponding to different moving ions (the $6M + 1$ scaling). The code currently implements the first level of parallelization and scales in a similar fashion to spICED. Implementation of the second level of parallelization is currently underway and will allow us to treat large numbers of moving CEID ions: currently we are limited to the treatment a maximum of 40 CEID ions, depending on the size of the system under investigation. As an example, the treatment of a large number of CEID ions (>100) is crucial in order to correctly describe electron-phonon couplings in photoexcited conjugated polymers and in the subsequent investigation of the influence of these couplings on the dynamics.

Very recently, a revised formalism for CEID has been developed to address some theoretical and numerical features of the original

Continued overleaf.

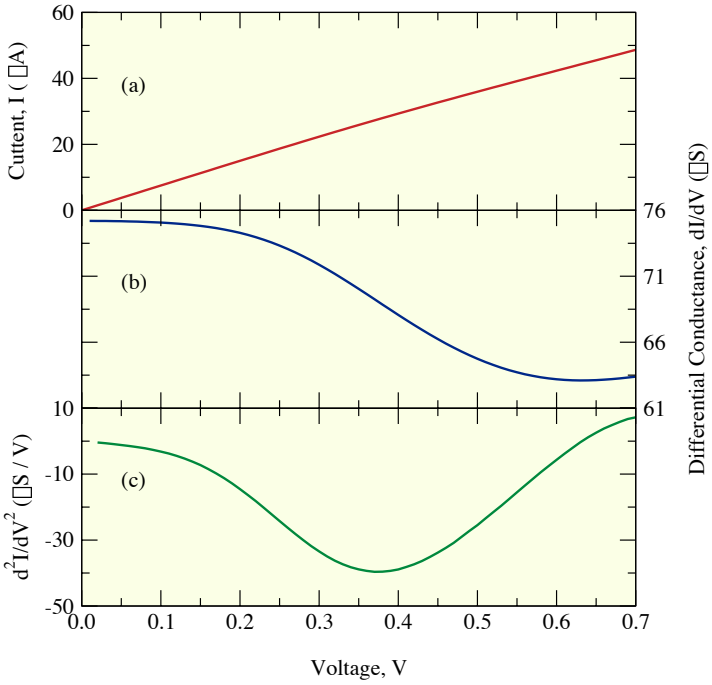


Figure 4: current-voltage (I-V) spectrum of a perfect chain. The chain consists of 601 gold atoms (with the atomic masses set to one atomic mass unit), the three central atoms being allowed to move. The equilibrium bond length of all atoms in the chain is 2.45Å. In the Figure, (a) represents the current as a function of applied voltage, (b) is the differential conductance and (c) the derivative of the differential conductance.

Modelling non-adiabatic processes continued

truncation scheme [4]. Numerical simulations of a model two-level system based on this new approach have provided further evidence of the soundness and generality of the CEID idea. As a consequence, this formalism – originally introduced for purely theoretical reasons – has made necessary substantial code development at UCL. Simulations of test models involving many degrees of freedom (e.g. the spin-boson model) are now under consideration at UCL to validate the scheme for more complex and physically relevant cases. Once again, the long-term goal of this revised CEID formalism is a general-purpose code to simulate the non-adiabatic dynamics of a many-atom system, and in particular non-radiative processes in polymers. A final merging of this code into pDINAMO is also an interesting future possibility.

To date, pDINAMO has been extensively applied to the study of electrical conduction in atomic wires at QUB. Experimentally, it has become possible to make chains of single atoms between two electrodes, pass a current along this chain and study the effects of the current flow on the atoms in the wire. A typical setup is shown in Figure 3. Theoretically, it has become possible to calculate the current that flows through an atomic wire and to make predictions about the local heating of the wire, about forces that the current exerts on individual atoms, and about the critical voltages at which the current blows the wire to pieces [5]. These predictions have been verified experimentally [6]. Most recently, a fresh wave of theory has started to focus on the dynamical quantum description, in real time, of current and of the interaction between current-carrying electrons and vibrating ions in nanodevices. In addition to current-induced local heating, we are interested in modelling inelastic current-voltage (I-V) spectroscopy [7]. Correlated inelastic interactions between electrons and ions under current flow result in inelastic features in the I-V spectrum, which appear as peaks or dips in the second derivative of current with respect to voltage. Figure 4 illustrates the I-V spectrum of a perfect chain, with equilibrium bond length of 2.45Å. In the simulation, the chain consists of 601 gold atoms (with the atomic masses set to one atomic mass unit), the three central atoms being allowed to move. The I-V features were obtained by calculating the steady-state current for 100 applied voltages (100 separate calculations). In the absence of inelastic interactions, the current as a function of voltage would be a straight line with slope 77.46 μA/V (the Landauer conductance G_0). With the inclusion of electron-phonon interactions in this system, the current (Figure 4, a) is reduced due to inelastic backscattering and the conductance is no longer constant (as is seen from Figure 4, b and c). Further analysis of the time-dependent results has allowed us to ascertain that the dominant contribution to the inelastic spectrum comes from a vibrational mode corresponding predominantly to motion of the central atom with much smaller amplitudes on the other two. Future work in this area will include the simulation of more realistic geometries (going beyond simple atomic chains and considering much larger system sizes) and investigating the effect of electron-electron screening.

Acknowledgements

We would like to thank the following people: Cristián Sánchez, Universidad Nacional de Córdoba, Argentina, the original author of DINAMO and Matt Harvey at Imperial College HPC for help and advice developing spICED. This work is funded by EPSRC under grant numbers EP/C006739/1, EP/C524403/1 and EP/C524381/1.

References

[1] Horsfield A P, Bowler D R, Fisher A J, Todorov T N and Sánchez C G, J Phys: Cond Matt 16 8251 (2004)

[2] Horsfield A P, Bowler D R, Fisher A J, Todorov T N and Sánchez C G, J Phys: Cond Matt 17 4793 (2005)

[3] Mason D R, le Page J, Race C P, Foulkes W M C, Finnis M W and Sutton A P, J Phys: Cond Matt 19 436209 (2007)

[4] Stella L, Meister M, Fisher A J and Horsfield A P, Accepted for publication in J Chem Phys (2007)

[5] Todorov T N, Hoesktra J and Sutton A P, Phys Rev Lett 86 3606 (2001)

[6] Tsutsui M, Kurokawa S and Sakai A, Nanotechnology 17 5334 (2006)

[7] McEniry E J, Bowler D R, Dundas D, Horsfield A P, Sánchez C G and Todorov T N, J Phys: Cond Matt 19 196201 (2007)

The MSc in HPC: looking back, looking forward

Judy Hardy, EPCC

At the time of writing, the sixth year of the MSc in HPC has just finished and we are now welcoming the new intake of students starting this September. Over the past six years, about seventy students have successfully completed the programme. Many of our graduates are now well-established in a range of software engineering or academic careers – including several who are members of staff at EPCC!

The dissertation is an important part of the MSc. Most students find the opportunity to work on an individual research project very rewarding and enjoyable. A wide variety of research topics are undertaken each year; this reflects the diverse range of software development and HPC-related projects within EPCC. Most of the dissertation reports are available on the MSc website. Here, we highlight two dissertations from 2006/07 to give a taste of the range of work carried out by our MSc students.

For further information about the MSc in HPC see http://www.epcc.ed.ac.uk/msc or email msc@epcc.ed.ac.uk

Optimizing parallel 3D Fast Fourier Transformations for a cluster of IBM POWER5 SMP nodes

Ulrich Sigrist (supervisor Joachim Hein)

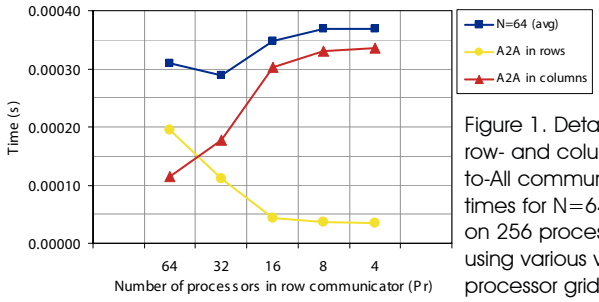


Figure 1. Details of row- and column All-to-All communication times for N=64 on 256 processors using various virtual processor grid layouts.

The Fast Fourier Transformation (FFT) of three-dimensional data is of particular importance for many numerical simulations used in High Performance Computing codes. We investigate how to optimize the parallel 3D FFT using a two-dimensional (2D) data decomposition on HPCx. We examined the properties of the IBM HPS interconnect with respect to the All-to-All communication pattern which is crucial for the parallel FFT. We investigate how the mapping of the virtual 2D processor grid to the processors in the SMP nodes affects the overall performance of the parallel 3D FFT. Furthermore, we compare the performance of the two different MPI All-to-All communication routines and the impact of using derived data types compared to manual buffer packing.

We have found that we can improve the overall performance by selecting an appropriate processor mapping. The optimal mapping depends on the problem size and the number of available processors. Mappings which only use communications within one SMP node for the All-to-All within the rows of the virtual 2D processor grid, result for most cases in lower total communication times than mappings which involve the network for both All-to-All steps. We found that mappings which result in network communications for both All-to-All steps can be beneficial if we process relatively small problem sizes on large numbers of processors. We conclude that, for small send sizes up to 2 KB, MPI_Alltoall is the faster than MPI_Alltoallv for all configurations which we examined. We have also found that using derived data types for the All-to-All communication can yield better performance compared to the manual buffer packing and unpacking.

Performance analysis and optimisation of LAMMPS on XCmaster, HPCx and Blue Gene/L

Geraldine McKenna (supervisors Fiona Reid, Michelle Weiland)

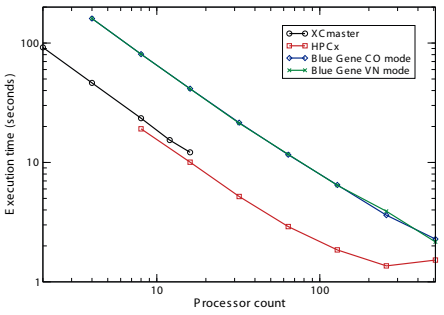
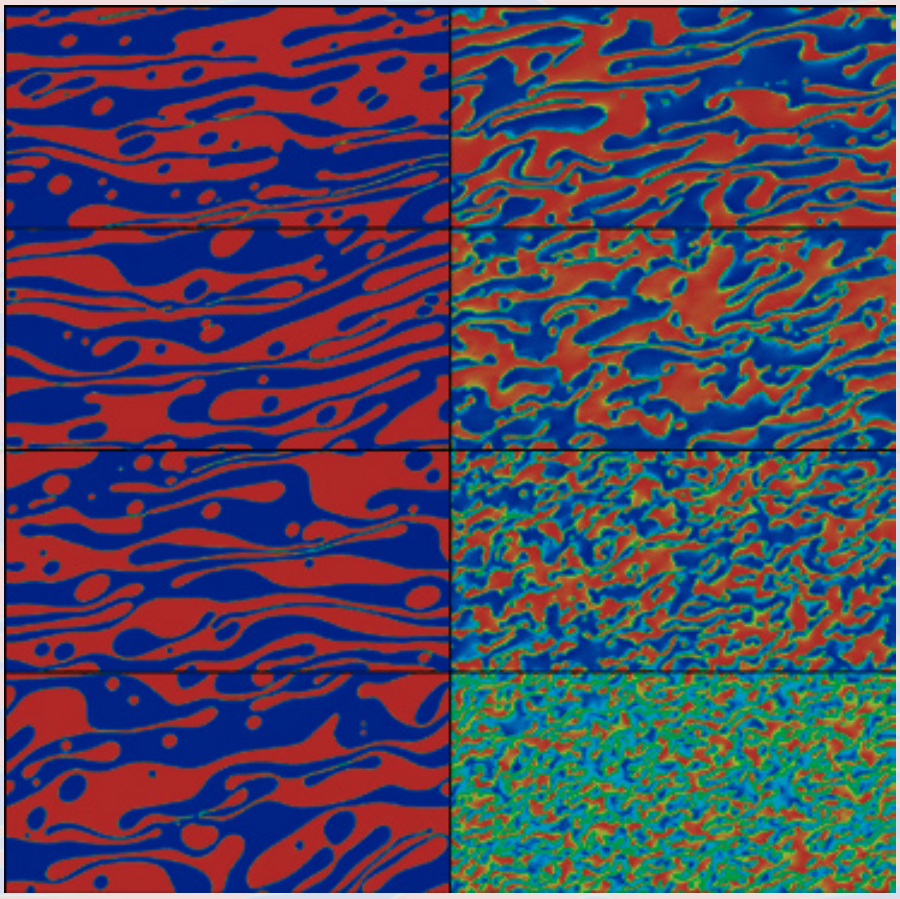


Figure 1: Comparison of the execution time of LAMMPS using the 2048000 atom rhodopsin protein benchmark on XCmaster, HPCx and Blue Gene (CO and VN modes).

This project focused on investigating the performance of LAMMPS molecular dynamics code on three different systems: a Hewlett Packard Itanium cluster based at Queens University Belfast (xcmaster), HPCx, and EPCC's single frame IBM Blue Gene/L. The code has been successfully ported to each of the systems and the speedup of LAMMPS is investigated on up to 1024 processors on HPCx, the Blue Gene, and up to 16 processors on xcmaster. The code is found to scale well to 256 processors on the HPCx system.

Profiling was used to identify the computationally expensive components of the code. We identified four main candidates for optimisation, and each of these were individually examined and a number of optimisation strategies were tested. Optimising these components improved the parallel performance of LAMMPS up to 18% depending upon the benchmark used.

The effect of simultaneous multithreading (SMT) on HPCx was also investigated to determine any potential performance improvement. The LAMMPS code was benchmarked with and without SMT enabled. We found that enabling simultaneous multithreading on HPCx can give a significant improvement in runtime without any changes to the source code, especially at lower processor counts. But as usual with many compute intensive applications, at higher processor counts SMT doesn't always increase performance.



The figures show two-dimensional lattice Boltzmann simulations of binary fluids (the two colours represent the two fluids) driven by shear flow. The different panels represent different Reynolds numbers, which can be varied over a wide range. The algorithm is highly parallel and scales well to large machines.

Programming, parallelism, petaflops... panic!

Kevin Stratford, EPCC, University of Edinburgh, UK

Moore and more...

Recent advertising campaigns by large microprocessor manufacturers such as Intel and AMD have concentrated on the benefits of their latest dual-core and multi-core offerings. The move to 'multi-core' technology is one response of the manufacturers to the increasingly difficulty of producing single processors with ever-higher clock rates.

Increases in processor speed have driven remarkable performance improvements over the last few decades, a fact encapsulated in a statement by Gordon Moore in 1965 that has become known as Moore's Law (Interestingly, Gordon Moore recently re-echoed his 1965 remarks by saying that engineering progress in reducing individual transistor size should still be possible for another decade). However, one factor mitigating against higher clock rates is the associated increase in power consumption, and hence cost of operation. The cost of electricity and cooling in large data centres and computer rooms is now a serious economic issue. As a direct consequence, high performance computing has already seen a move to larger numbers of less power-hungry processors, for example the IBM Blue Gene L and P series machines. The resurrection the somewhat arcane term 'core' has thrown the lexicon of processing elements into confusion – is it nodes, processors, cores, chips or CPUs anyone? The more serious question for application programmers is: how does one program these more complicated hardware architectures efficiently?

Proponents of parallel computing will argue that they have been banging on about the solution for years. Two standard techniques are in regular use for writing parallel programs. There are OpenMP, a directive-based approach for shared memory machines which have a global address space, and the message passing interface (MPI) which is used to provide a standard way to exchange data between processors which do not share memory. Of these OpenMP is generally regarded as the more straightforward, while MPI is more portable because it does not require shared memory hardware. Even with these standard approaches, one often hears calls for more simple ways to program in parallel. There are a number of candidates, eg Unified Parallel C and Co-Array Fortran language extensions, which allow programmers to more easily implement parallel algorithms. High Performance Fortran has also been used successfully on a number of platforms. However, the application programmer wishing to write portable and long-lasting code is left with a slightly uneasy feeling that these programming models may not continue to be – or ever be – implemented. The hope is that with the move to commodity machines using parallelism, tools for application programmers will only improve.

Continued opposite.

HPC-Europa: building bridges in European computational science

Catherine Inglis, EPCC

HPC-Europa's Transnational Access programme allows European computational scientists to travel abroad to collaborate with another research group in a similar field[1], while also gaining access to some of the most powerful computing facilities in Europe. The programme is fully funded by the EC, and travel and living costs are provided.

Since the programme began in 2004, HPC-Europa has approved 702 applications from researchers working in 29 European countries[2]. These visitors have come from a variety of disciplines, ranging from traditional computationally-demanding fields, such as computational physics or chemistry, to emerging disciplines, such as bioinformatics, and even including human sciences, such as one project on "Linguistic and stylistic analyses of Greek and Latin texts by computer".

The programme is open to researchers of any level, with a roughly equal mix of postgraduate students, post-docs and experienced researchers taking part.

The first cycle of funding comes to an end in December 2007, and we are currently in the Contract Negotiation phase for a one-year extension to the project. Early next year we hope to submit a proposal for a further five years of funding under the EC's Framework Seven funding. Please keep an eye on the HPC-Europa website for further updates.

Readers of Capability Computing may wish to apply to visit one of the research groups associated with the programme, or alternatively you may wish to host a visitor within your own research group. Please see our web pages for further information, including details of the on-line application procedure, or contact access@hpc-europa.org if you have any questions which are not answered there.

http://www.hpc-europa.org/

1. Research groups must be associated with one of the participating centres: BSC (Barcelona), CINECA (Bologna), EPCC (Edinburgh), HLRS (Stuttgart), IDRIS (Paris), SARA (Amsterdam). See the HPC-Europa website for further information.

2. To be eligible to apply for the programme, researchers must currently be working in one of the following countries: any of the 27 EU member countries, Iceland, Israel, Liechtenstein, Norway, Switzerland or Turkey. Researchers may not visit another research group in their own country through this programme.

Even then, the problem of how to program efficiently very large numbers of processors is one which is only just beginning to be addressed. The next generation of petascale machines are likely to have hundreds of thousands or even millions of processors. Whether the traditional methods such as MPI will stand up in this new regime is yet to be seen. Such large numbers of processors might make it impractical to decompose a single problem, and keep the overheads of load balance and communication under control. These problems will encourage the use of fundamentally parallel algorithms, and manageably parallel ensemble methods (which are increasingly important in areas like weather forecasting).

Do not panic

Improvements to algorithms can arise when coming at a familiar problem from a different perspective. One example is the lattice Boltzmann equation, which is used by a number of groups in the UK, eg RealityGrid [1] and elsewhere, particularly those with a statistical mechanics background. Here, instead of solving directly the Navier-Stokes equations, one solves a discrete Boltzmann equation which approximates the Navier-Stokes equations in the limit of low Mach number (the ratio of the characteristic flow velocity to the speed of sound in the fluid). Traditional methods for incompressible Navier-Stokes flows often require the solution of a global problem for the fluid pressure at each time step, reflecting that sound waves propagate essentially instantaneously through the fluid. A number of methods exist to tackle this sort of problem in parallel, including multigrid and pseudospectral methods. The latter, for example, relies on Fast Fourier Transforms and so can scale reasonably well with the system size. Ultimately, however, the global nature of the problem is a barrier to parallelisation.

The trick in lattice Boltzmann is to relax the exact incompressibility constraint so that the speed of sound is finite, and the pressure calculation becomes local. This means the calculation can be scaled efficiently to extremely large numbers of processors. Lattice Boltzmann also has the curious capability to manipulate its internal set of units so that appropriate control parameters (for example, length and time scales) can be varied over many orders of magnitude with some ease. This may be of increasing importance as 'multi-scale' physics problems become more prevalent. Such rethinking could yield local parallel algorithms for problems which are apparently global.

Wherever improvements come from, it is sure that they will (in the words of the advertising industry) 'unleash the power' of more powerful multi-core or multi-processor machines.

[1] RealityGrid: http://www.realitygrid.org/

Service Administration from EPCC (SAFE)

Stephen Booth, EPCC

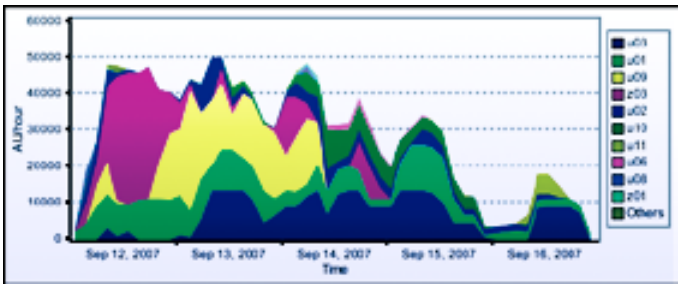


Figure 1. Use by project for the first 5 days of the HECToR reliability trial, produced by the SAFE.

Figure 2. Report generation in the SAFE.

EPCC is involved in providing supercomputing services on a wide variety of systems. These include two national supercomputing services (HPCx at the STFC Daresbury laboratories, and HECToR at the University of Edinburgh). In addition we also run services on local systems such as EPCC's IBM Blue Gene/L and our FPGA system, Maxwell. SAFE (Service Administration from EPCC) is the software system we use to help provide these services.

The aim of the SAFE system is to provide a single unified administration system tailored to the needs of a supercomputing service but flexible enough to support different types of computer system with different operating systems and running different software. The responsibilities of the SAFE system include:

- User Registration
- Accounting
- Resource Management
- Report Generation
- Contact Management
- Helpdesk and Support

Though many of these functions are common to many other types of computer support operation, a supercomputing service is distinct in many ways. Supercomputing services draw their users from many different organisations and geographical locations. This means that the SAFE has to be an entirely stand-alone system and cannot leverage off existing systems such as University account registration procedures. Because supercomputers are shared resources, accounting, resource management and report generation are of particular importance. Even providing a helpdesk for this kind of service has additional challenges. Many of the users of a supercomputing service are programmers who develop their own application codes rather than using standard applications. This means the helpdesk has to have particularly good support for exchanging source code and data files. Users can submit helpdesk requests via email or directly via the SAFE web interface, and both of these routes provide full support for MIME attachments. Users can view the progress of their request via the web interface, including downloading any attachments in the request log. Standard applications are also important, and there are often a large number of these provided in a variety of different ways by different organisations. The helpdesk has to be able to assign problem requests to external organisations if necessary.

The SAFE is a web application built using Java Servlets and JSP pages, and uses a MYSQL database to hold its information. This makes it highly portable and it can be hosted on any system where Java is available. Users interact with the SAFE using web browsers and by email.

Within the SAFE, different types of resource are allocated to projects. Each project has a number of designated managers, and where possible, decisions about resource management within a project are devolved to these project managers. Standard users apply for accounts and register with the SAFE via the web. Once this request has been approved by their project manager, the SAFE generates change request tickets asking for the account to be set up. Project managers can also use the SAFE to allocate computer time budgets for any subprojects they define, and set disk quotas if desired. Accounting and disk quota information is automatically uploaded into the SAFE, allowing users to view the current state of their accounts and to produce detailed charts and reports about their usage.

ISC2007

Dresden, Germany

Chris Johnson and Kenton D'Mellow, EPCC

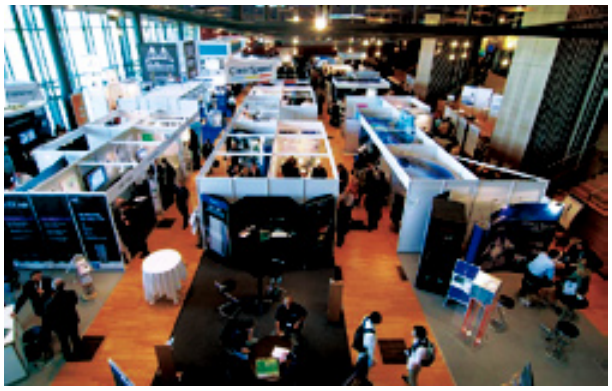


Image courtesy of ISC2007.

This year marked the return of the International Supercomputing Conference [1] to Dresden after its very successful debut there in 2006. Dresden's beautifully restored city centre boasts an impressive new conference centre, facing onto the calm waters of the River Elbe.

Yet again, attendance at ISC2007 broke previous records with over 1200 delegates and 85 exhibitors, increasingly affirming this event's importance on the world stage. The busy and dynamic atmosphere at the conference reflects directly that of the current HPC market: the breadth of technologies showcased this year was wider than ever. A strong undercurrent to this year's event was the mass movement to multi-core and so-called 'many-core' technologies, in an effort to stay ahead of Moore's Law while reducing power consumption. It was clear, from both the exhibition and conference sessions, that novel architecture technologies such as heterogeneous multi- and many-core CPU's, GPGPUs, hardware accelerators, ASICs and FPGAs are becoming key components in modern HPC solutions in a continual drive to give a performance edge at low cost. Currently HPC seems to have a plethora of options.

A central part of ISC is the announcement of the TOP500 list of the world's fastest supercomputers [2]. In the 29th TOP500 list, there were no changes to the number 1 spot: for the fourth time in a row, the large IBM Blue Gene/L system at Laurence Livermore National Laboratory sits comfortably in the lead with a 280.6 Tflop/s Linpack benchmark performance – head and shoulders above the rest. However, with several Petaflop-scale roadmaps being announced over the last year or so, it will be interesting to see if the Blue Gene/L remains there much longer.

[1] http://www.supercomp.de/

[2] http://www.top500.org/

Image courtesy of Joachim Hein.

SCICOMP 2007

Garching, Germany

Joachim Hein, EPCC

The 13th meeting of ScicomP, the IBM System Scientific Computing User Group, was held in Garching from the 16th to the 20th July 2007. The meeting was hosted by the Rechenzentrum Garching (RZG) of the Max Plank Society and the Max Planck Institute for Plasma Physics. The summer meeting of SP-XXL was co-located with ScicomP. In contrast to ScicomP, which focuses on the application aspects of scientific computing, SP-XXL has a stronger interest in the systems software and the operational aspects of large IBM compute systems.

The ScicomP program offered an attractive mix of presentations. IBM staff members described the latest products from IBM. This included POWER6 and Cell processor based compute platforms, as well as the new Blue Gene/P. On the systems software side, key presentations focused on the latest additions and improvements to the IBM XL compiler suite – the latest tools from IBM's ACTC group and the parallel environment (PE), which contains the MPI library. These products are highly relevant to the HPCx user community.

In a second set of presentations, users from computer centres, research institutions and universities reported on their experiences. A particular highlight was the keynote presentation by Wolfgang Hillebrandt from the Max Plank Institute for Astrophysics in Garching. He spoke about large eddy simulations in thermonuclear flames and their relevance to Type Ia Supernovae. As part of the DEISA Extreme Computing Initiative (DECI), his project had access to HPCx and used a substantial amount of compute resources.

From the HPCx team, Mark Bull (EPCC) and Michal Piotrowski (EPCC) both contributed talks. Mark Bull reported on his investigation of a wide range of HPCx applications using the hardware counters of the POWER5 chip. For details, see his article in this issue and his technical report HPCxTR0703 available on the HPCx website. Michal Piotrowski discussed mixed mode programming on HPCx. His presentation is based on his final project on EPCC's MSc in HPC. Michal's dissertation is available on EPCC's website and an HPCx technical report will be available in the near future. The slides for all presentations given at the meeting are available from the ScicomP website.

Attendees had the opportunity to visit the new machine building at the Leibniz Rechenzentrum (LRZ). The HLRB II machine, based on SGI Altix hardware, is their current flagship service. HLRB II offers a sustained performance of 57 TFlop/s on Linpack and is presently ranked as number 10 on the TOP500 list. The ScicomP program also included a visit to the Nymphenburg Palace and the 'Chinesischer Turm' in Munich's English Garden.

18TH DARESBURY MACHINE EVALUATION WORKSHOP
HOLIDAY INN RUNCORN, 27-28 NOVEMBER 2007

PRESENTATIONS
Talks from national and international HPC experts and vendors.

EXHIBITION
Vendor exhibits and benchmarking facilities - servers, clusters, storage, visualization...

Try out your own codes on different machines, which will be accessible over the internet prior to the workshop, as well as during the event.

FURTHER DETAILS
http://www.cse.scitech.ac.uk/disco/mew18

Machine_Evaluation_Workshop@dl.ac.uk

+44 (0)1925 603240 / +44 (0)1925 603805

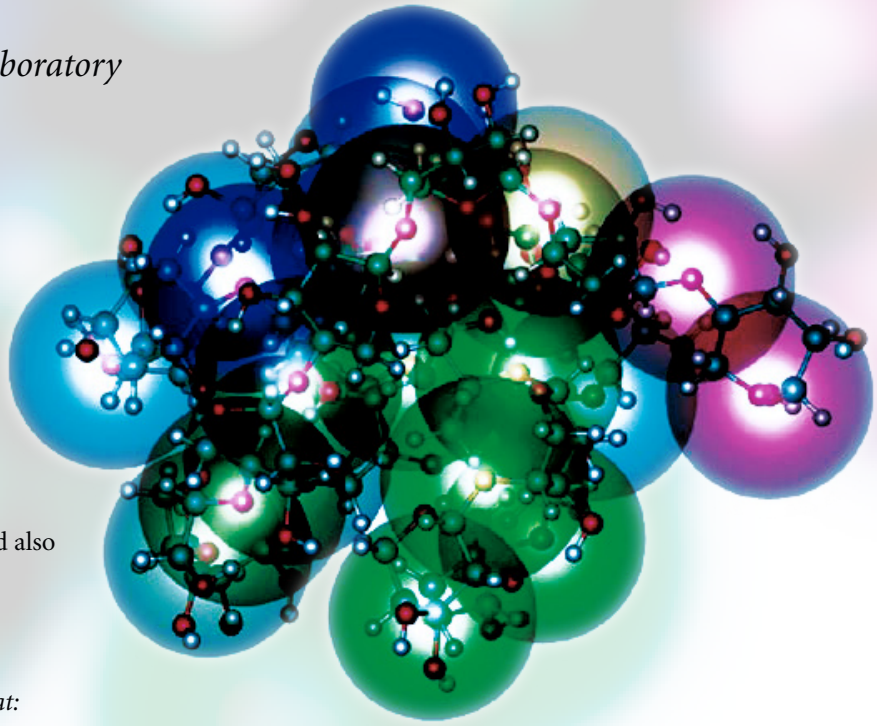
Supercomputing 2007
12–16 November, Reno, NV, USA

SC07

EPCC will exhibit in booth 3019. STFC Daresbury Laboratory will be in booth 3015.

Fifth HPCx Annual Seminar

26th November 2007, STFC Daresbury Laboratory



The fifth HPCx Annual Seminar will be held in the Merrison Lecture Theatre, STFC Daresbury Laboratory on Monday 26th November 2007.

The seminar will include presentations from HPCx users and staff. It is the latest in a highly successful series of events organised by the UK's primary high performance computing service.

The event will focus both on the current use of HPCx and also the new HECToR service.

Attendance at this event is free for all academics.

Further details and registration information can be found at:
http://www.hpcx.ac.uk/about/events/annual2007/